

Data Analysis of Drugs, Side Effects, and Medical Conditions

A Data Analytics Project Report

Submitted by:- DALWADI MEET

Enrollment /UNID :- UMID12112570217

Internship: - Data Analytics

Submitted to:-

UNIFIED MENTOR

Tools Used: -

Python (Jupyter Notebook), SQL, Power BI, Microsoft Excel

Internship Year :-

2025 – 2026

Date of Submission :-

15 / 12 / 2025

1. Introduction

In the healthcare and pharmaceutical industry, understanding how drugs perform in real-world usage is extremely important. Drugs are prescribed to treat various medical conditions, but they often come with side effects that can affect patient safety and satisfaction. Along with this, user ratings and reviews provide valuable feedback on how effective and tolerable a drug is for patients.

With the increasing availability of healthcare data, data analysis plays a crucial role in identifying patterns and relationships between drugs, their side effects, and the medical conditions they treat. Analyzing such data helps healthcare professionals, researchers, and pharmaceutical companies make better decisions regarding drug usage and safety.

This project focuses on performing **Exploratory Data Analysis (EDA)** on a drug dataset that contains information about drugs, their medical conditions, side effects, safety indicators, ratings, and reviews. Using tools like **Python (Jupyter Notebook)** for data analysis and **Power BI** for visualization, the project aims to uncover meaningful insights and present them in an interactive and easy-to-understand manner.

2. Objective of the Project

The objective of this project is to perform an in-depth analysis of pharmaceutical drug data in order to understand the relationships between drugs, their side effects, and the medical conditions they are prescribed to treat. In addition, the project aims to examine how patient ratings and reviews reflect the effectiveness, safety, and overall user experience of these drugs.

Drugs used to treat medical conditions often produce side effects that may impact patient satisfaction and adherence to treatment. By analyzing side effects along with drug ratings and review counts, this project seeks to identify patterns that can help assess drug performance and potential risk factors. Understanding these relationships is important for healthcare professionals, researchers, and pharmaceutical organizations when evaluating drug safety and effectiveness.

Another key objective of this project is to explore how drug characteristics such as drug class, prescription status (Rx or OTC), pregnancy risk category, alcohol interaction, and controlled substance classification influence user ratings and reported side effects. Comparing these attributes helps highlight differences in safety and patient response across various categories of drugs.

The project also focuses on identifying the most common medical conditions treated by drugs and analyzing whether certain conditions are associated with higher side-effect frequency or lower user ratings. This analysis helps in understanding which conditions may require closer monitoring or alternative treatment approaches.

Finally, the project aims to present all findings through interactive and visually appealing dashboards using Power BI. These dashboards allow users to filter data by medical condition, drug class, and safety indicators, enabling deeper exploration of drug-side effect relationships. Overall, the objective is to transform raw drug data into meaningful insights that support better understanding of drug safety, effectiveness, and patient experience.

3. Dataset Description

The dataset used in this project contains detailed information about various pharmaceutical drugs that are prescribed or sold to treat a wide range of medical conditions such as acne, infections, cardiovascular diseases, neurological disorders, mental health conditions, and chronic illnesses. The dataset was collected from publicly available drug information sources and includes both medical and user-generated data such as ratings and reviews.

This dataset provides a comprehensive view of drugs by combining their medical properties, safety indicators, and patient feedback. Such a combination makes the dataset suitable for exploratory data analysis aimed at understanding drug effectiveness, safety concerns, and patient satisfaction.

3.1 Structure of the Dataset

Each row in the dataset represents a **single drug**, while the columns describe various attributes related to that drug. The dataset includes the following important columns:

Drug Identification Columns

- **drug_name:** Represents the commonly used name of the drug. This column is essential for identifying and comparing individual drugs.
- **generic_name:** Represents the chemical or generic name of the drug. Generic names help group drugs with similar compositions but sold under different brand names.
- **brand_names:** Contains the commercial brand names under which the drug is available in the market.

Medical Condition Information

- **medical_condition:** Specifies the medical condition or disease that the drug is used to treat, such as pain, hypertension, acne, or anxiety.
- **medical_condition_description:** Provides a brief explanation of the medical condition, which helps in understanding the context of drug usage.
- **medical_condition_url:** Contains a URL linking to additional information about the medical condition.

These columns are useful for analyzing how drugs are distributed across different medical conditions and identifying which conditions have more treatment options available.

Drug Classification and Usage

- **drug_classes:** Indicates the class to which the drug belongs, such as antibiotics, antidepressants, or antihistamines. Drugs within the same class typically share a similar mechanism of action.
- **activity:** Represents the activity status of the drug based on recent user engagement and availability.

- **rx_otc:** Indicates whether the drug requires a prescription (Rx), is available over the counter (OTC), or falls under both categories.

These attributes help compare prescription drugs with over-the-counter medications and analyze differences in ratings and side effects.

Safety and Risk Indicators

- **pregnancy_category:** Classifies drugs based on their safety during pregnancy (A, B, C, D, X, or N). This column is crucial for understanding potential fetal risk.
- **csa:** Represents the Controlled Substances Act (CSA) schedule, indicating the drug's potential for abuse or dependence.
- **alcohol:** Indicates whether the drug interacts with alcohol, marked by an "X" for interaction.

These safety indicators are essential for analyzing drug risks and understanding how safety concerns influence user ratings and reviews.

Side Effects Information

- **side_effects:** Contains textual information describing common side effects associated with the drug. This column often includes multiple side effects listed in a single string, separated by commas or semicolons.
- **related_drugs:** Lists other drugs related to the primary drug, which can be useful for comparative analysis.

The side effects column plays a central role in this project, as it is used to analyze the frequency, type, and impact of side effects on drug ratings.

User Feedback

- **rating:** Represents the average user rating of the drug on a scale of 1 to 10, where higher values indicate better effectiveness and user satisfaction.
- **no_of_reviews:** Indicates the number of user reviews submitted for the drug.

These columns provide insight into patient experience and are used to evaluate how side effects and safety indicators affect user perception.

Reference Information

- **drug_link:** Contains a URL linking to detailed information about the drug.

4. Methodology and Data Cleaning

This project follows a structured methodology to analyze drug-related data and extract meaningful insights. The overall approach consists of data collection, data cleaning, exploratory data analysis, and visualization. Python (Jupyter Notebook) was used for data preprocessing and analysis, while Power BI was used for creating interactive dashboards.

4.1 Methodology Overview

The methodology adopted in this project includes the following steps:

1. Loading the raw dataset into Jupyter Notebook
2. Understanding the data structure and identifying data quality issues
3. Cleaning and preprocessing the data
4. Performing exploratory data analysis (EDA)
5. Visualizing results using Power BI

Each step was carefully executed to ensure accurate and reliable analysis results.

4.2 Data Loading

The dataset was loaded into Jupyter Notebook using the pandas library. Basic inspection was performed to understand the size, structure, and data types of the dataset. Functions such as `.head()`, `.info()`, and `.describe()` were used to examine the data.

4.3 Handling Missing Values

Several columns in the dataset contained missing or null values. The following strategies were applied:

- **Numerical columns** such as rating and no_of_reviews were analyzed for missing values. Rows with missing ratings were excluded from rating-based analysis to avoid misleading results.
- **Categorical columns** such as side_effects, brand_names, and alcohol were cleaned by removing or filtering null values where required.
- Missing values that did not significantly affect the analysis were handled using appropriate filtering techniques.

This approach ensured that missing data did not distort analytical outcomes.

4.4 Cleaning and Processing Side Effects Data

The side_effects column contained multiple side effects in a single text field, separated by commas or semicolons. To enable meaningful analysis:

- The column was split using delimiters (, and ;)
- Side effects were transformed into individual rows
- Extra spaces were removed and text was standardized

This normalization allowed accurate counting and comparison of individual side effects across drugs.

4.5 Data Type Conversion

Certain columns required data type conversion for analysis:

- Ratings were converted to numeric values to allow aggregation and comparison
 - Review counts were converted to integer values
 - Categorical fields such as pregnancy category and CSA schedule were standardized for consistency
-

4.6 Feature Engineering

New derived fields were created to support analysis:

- A flag indicating alcohol interaction (Yes/No)

	alcohol_yes	alcohol_no
	bigint	bigint
1	1377	1554

- Categorized risk levels based on CSA schedules

	csa	total_drugs
	text	bigint
1	N	2688
2	2	101
3	4	71
4	3	26
5	5	20
6	M	16
7	U	9

These derived features enhanced the depth of analysis and helped in identifying relationships between safety indicators and user feedback.

4.7 Data Validation

After cleaning and preprocessing, the dataset was validated to ensure:

- No duplicate records were present
- Side effects were properly normalized
- Top reviews Drugs

	drug_name text	medical_condition text	no_of_reviews bigint
1	phentermine	Weight Loss	2934
2	bupropion / naltrexone	Weight Loss	2013
3	Contrave	Weight Loss	1939
4	escitalopram	Anxiety	1471
5	Saxenda	Weight Loss	1377
6	bisacodyl	Constipation	1357
7	nitrofurantoin	UTI	1242
8	buspirone	Anxiety	1034
9	zolpidem	Insomnia	1008
10	isotretinoin	Acne	999

This validation step ensured the reliability of the dataset before visualization.

4.8 Data Visualization Approach

Cleaned and processed data was imported into Power BI. Interactive dashboards were created using charts, tables, and slicers to explore relationships between drugs, side effects, medical conditions, ratings, and reviews.

5. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was performed to understand patterns, relationships, and trends within the drug dataset. The analysis focused on examining how drugs are distributed across medical conditions, identifying commonly reported side effects, and understanding how these factors influence user ratings and reviews. Python (Jupyter Notebook) and Power BI were used to perform and visualize the analysis.

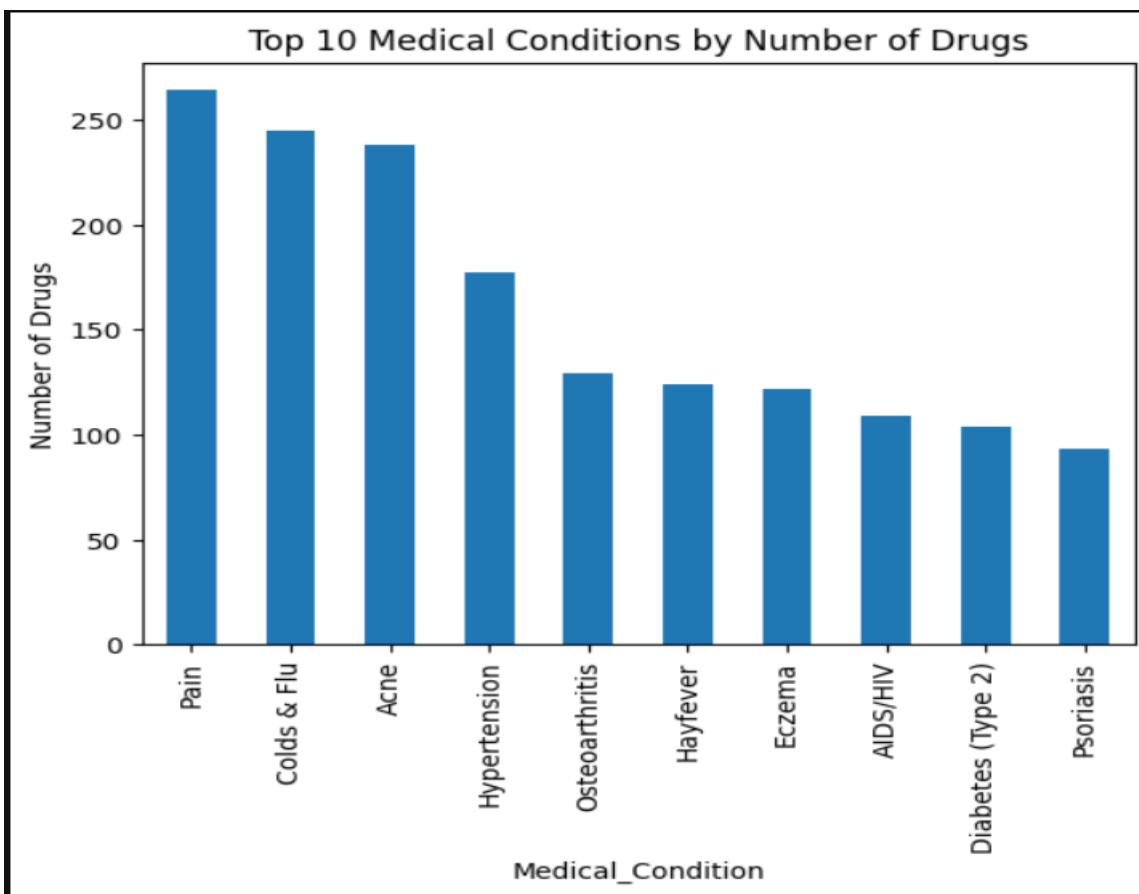
5.1 Analysis of Medical Conditions and Drug Distribution

The first step of the analysis involved examining how drugs are distributed across different medical conditions. The number of drugs available for each medical condition was calculated to identify conditions with the highest treatment availability.

The analysis revealed that certain medical conditions have a significantly higher number of associated drugs, indicating greater pharmaceutical focus and demand. Chronic conditions such as pain-related disorders, mental health conditions, and infections showed a higher concentration of drugs compared to less common conditions.

This analysis helps in understanding which medical conditions receive more pharmaceutical attention and may also reflect the prevalence of these conditions.

- Which medical conditions have the highest number of available drugs?



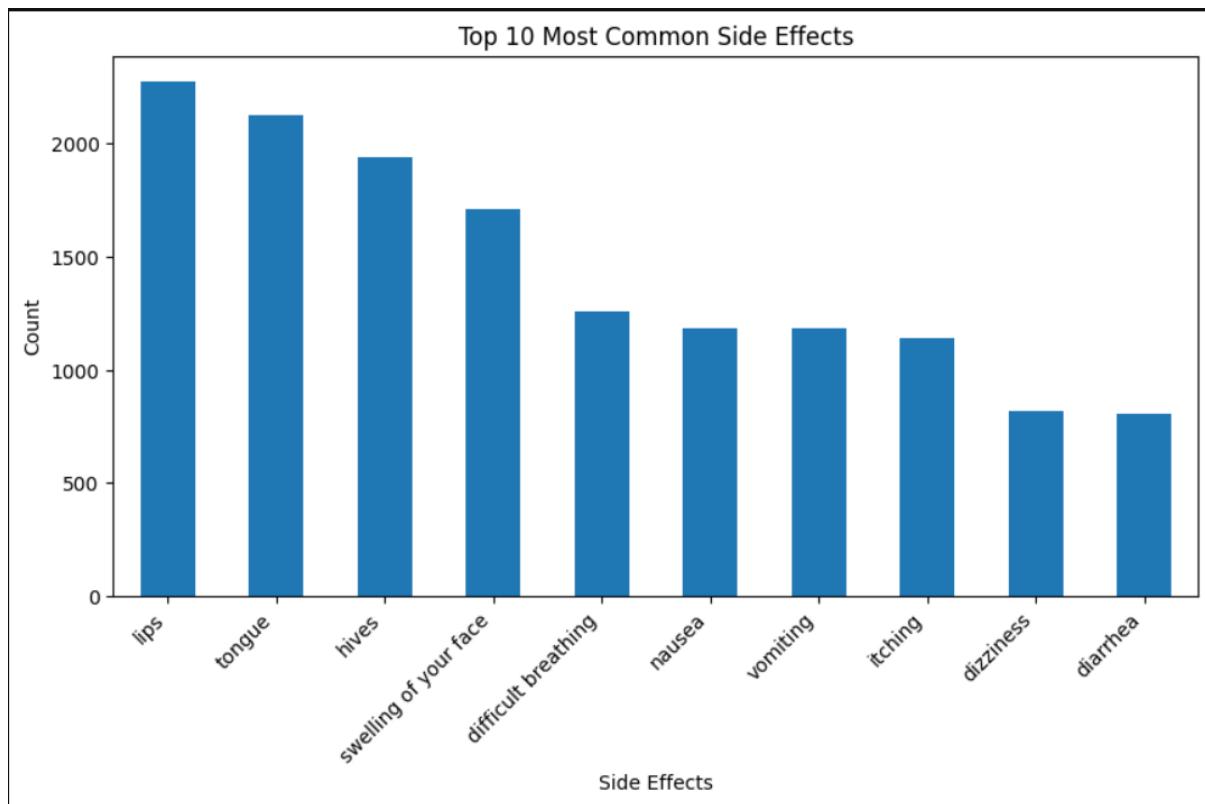
5.2 Side Effects Analysis

Side effects play a critical role in evaluating drug safety and patient satisfaction. The side_effects column was cleaned and normalized by splitting multiple side effects into individual entries. This allowed accurate frequency analysis of individual side effects.

The analysis identified the most commonly reported side effects across all drugs. Symptoms such as nausea, headache, dizziness, and fatigue were found to occur frequently across multiple drug classes and medical conditions.

This finding highlights that a small number of side effects are widely shared among many drugs, which may influence overall user experience and ratings.

- What are the most commonly reported side effects across all drugs?



5.3 Side Effects by Medical Condition

Further analysis explored the relationship between side effects and medical conditions. By grouping side effects by medical condition, it was observed that drugs used for chronic and long-term treatments tend to have a higher number of reported side effects.

This relationship suggests that patients undergoing long-term treatment may experience more adverse effects, which can impact adherence to medication and overall satisfaction.

	unique_medical_condition	lock
1		47

5.4 Drug Class and Side Effects Relationship

Drugs were grouped based on their drug classes to examine how side effects vary across different drug categories. Certain drug classes, such as antibiotics and antidepressants, showed a higher frequency of reported side effects compared to others.

This analysis helps identify drug classes that may require closer monitoring or more detailed patient guidance regarding potential side effects.

5.5 Ratings Distribution Analysis

User ratings were analyzed to understand how drugs are perceived in terms of effectiveness and ease of use. The distribution of ratings showed that most drugs received moderate to high ratings, with fewer drugs receiving extremely low scores.

The rating distribution provided a baseline for comparing how different factors such as side effects and safety indicators influence user feedback.

② What is the Avarage distribution of drug ratings?

②

```
; medical_condition
Herpes           7.690000
Anxiety          7.622222
Gout             7.577778
Erectile Dysfunction 7.530769
Swine Flu         6.920000
ADHD              6.805455
Depression        6.625490
Migraine          6.432787
Bipolar Disorder   6.248936
COPD              6.065217
Name: rating, dtype: float64
```

5.6 Impact of Side Effects on Ratings

One of the key objectives of the analysis was to examine whether side effects affect user ratings. The analysis showed that drugs associated with more frequent or severe side effects tend to have lower average ratings.

This negative relationship suggests that side effects play a significant role in shaping patient satisfaction and perceived drug effectiveness.

5.7 OTC/RX Analysis

How many drugs are prescription (Rx) compared to over-the-counter (OTC)?

	number_of_otc	number_of_rx	number_of_both
1	328	1998	604

5.8 Safety Indicators Analysis

Safety-related attributes such as alcohol interaction, pregnancy category, and Controlled Substances Act (CSA) schedules were also examined. Drugs with alcohol interaction warnings or higher CSA risk levels showed slightly lower average ratings compared to safer alternatives.

This suggests that safety concerns may influence patient perception and satisfaction.

	csa 🔒	total_drugs 🔒
1	N	2688
2	2	101
3	4	71
4	3	26
5	5	20
6	M	16
7	U	9

6. Dashboard Design and Visualization

6.1 Purpose of the Dashboard

The dashboard was designed to visually analyze and explore the relationships between **drugs, medical conditions, side effects, ratings, and reviews** in an interactive manner. The goal of the dashboard is to allow users to quickly identify patterns, compare drug performance, and assess safety-related factors such as alcohol interaction and prescription type.

Power BI was chosen as the visualization tool due to its ability to handle large datasets, provide interactive filtering, and present insights in a user-friendly format.

6.2 Dashboard Layout Overview

The dashboard follows a structured layout divided into four main sections:

1. **Filter Panel (Left Side)**
2. **Key Performance Indicators (Top Section)**
3. **Analytical Charts (Middle Section)**
4. **Detailed Drug Information Table (Bottom Section)**

This layout ensures clarity, usability, and logical data flow.

6.3 Filter Panel (Slicers)

The left-side panel contains slicers that allow users to dynamically filter the dashboard:

- **Medical Condition** – Filters drugs based on selected medical condition
- **Drug Name** – Allows selection of specific drugs
- **Drug Class** – Filters drugs by their drug classification
- **Rx / OTC** – Distinguishes between prescription and over-the-counter drugs

These slicers enable focused analysis. For example, selecting a specific medical condition updates all charts to display only drugs related to that condition.

6.4 Key Performance Indicators (KPIs)

The top section displays summary metrics that provide a quick overview of the dataset:

- **Total Medical Conditions** – Shows the number of unique medical conditions
- **Total Drugs** – Displays the total number of drugs available
- **Total Reviews** – Represents the sum of all user reviews
- **Average Rating** – Indicates overall user satisfaction

These KPIs help users understand the dataset scale at a glance.

6.5 Analytical Visualizations

6.5.1 Top Drugs by Reviews

- **Chart Type:** Horizontal Bar Chart
 - **Purpose:** Identifies drugs with the highest user engagement
 - **Insight:** Drugs with higher reviews provide more reliable feedback
-

6.5.2 Rating Distribution

- **Chart Type:** Column Chart (Rating Bins)
 - **Purpose:** Shows how drug ratings are distributed across values
 - **Insight:** Most drugs fall within mid-to-high rating ranges
-

These visualizations help explore how user feedback varies across drugs.

6.6 Detailed Drug Information Table

The table at the bottom provides detailed information for each drug, including:

- Drug name
- Medical condition
- Generic name
- Pregnancy category
- CSA schedule
- Rating

This table updates dynamically based on slicer selections, allowing users to drill down into specific drugs and compare safety and performance attributes.

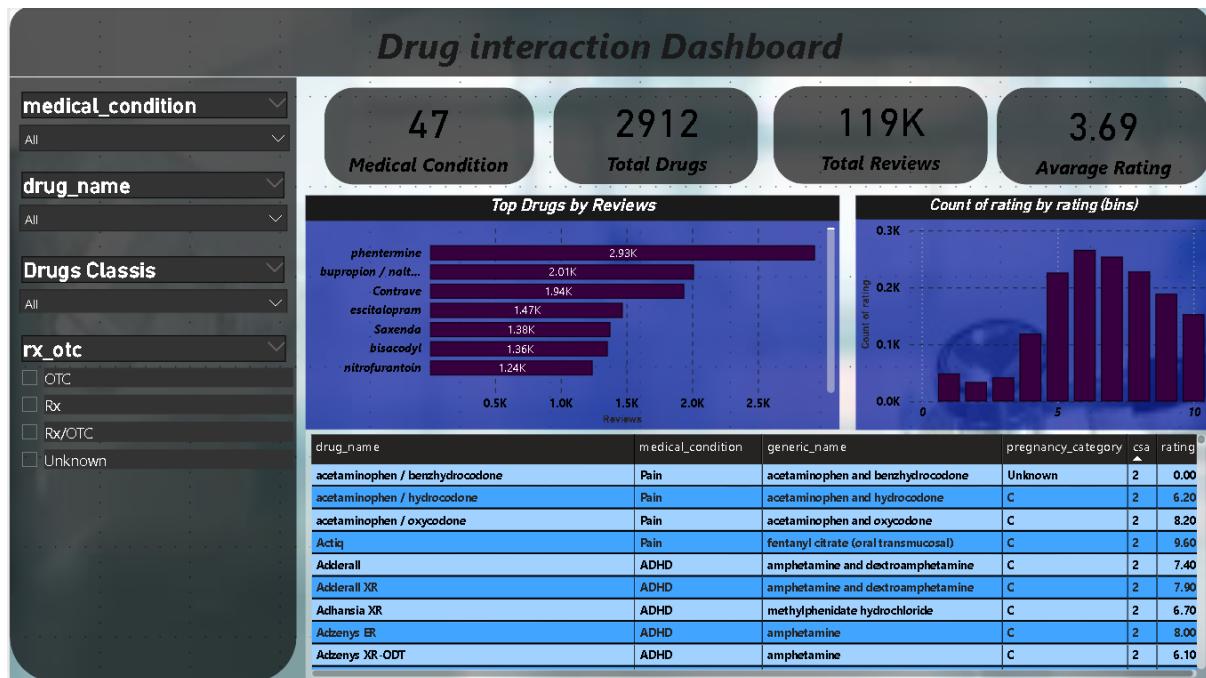
6.7 Interactivity and User Experience

All visuals are interconnected through Power BI's interaction features. Selecting a slicer or clicking on a chart element automatically updates all other visuals. This interactive capability allows users to explore complex relationships between drugs, side effects, and medical conditions without writing queries.

6.8 Alignment with Project Objective

The dashboard directly supports the project goal by:

- Linking drugs to medical conditions
- Highlighting user ratings and review patterns
- Allowing comparison between Rx and OTC drugs
- Enabling safety-focused analysis



7. Key Insights and Findings

This section summarizes the key insights obtained from the exploratory data analysis and the Power BI dashboard. The findings focus on the relationships between **drugs, medical conditions, side effects, ratings, and reviews**.

7.1 Drug Usage Across Medical Conditions

- A limited number of medical conditions account for a large portion of the drugs in the dataset.
- Common conditions such as **Pain, ADHD, Depression, and Anxiety** have the highest number of associated drugs.
- This indicates that chronic and widely prevalent conditions tend to have more treatment alternatives available in the pharmaceutical market.

7.2 Drug Popularity Based on Reviews

- Certain drugs receive significantly higher numbers of reviews compared to others.
- Drugs such as **phentermine, bupropion combinations, and escitalopram** show high user engagement.
- A higher number of reviews suggests wider usage and stronger patient feedback, making these drugs more reliable for analysis.

7.3 Ratings and User Satisfaction Trends

- The rating distribution shows that most drugs receive **moderate to high ratings**, typically between **6 and 9**.
 - Extremely low ratings are less common, indicating that most drugs provide acceptable effectiveness or tolerability.
 - Average rating values remain relatively stable across many drug classes, suggesting consistent perceived effectiveness.
-

7.4 Side Effects and Safety Observations

- Some side effects appear frequently across multiple drugs, indicating common reactions rather than drug-specific issues.
 - Drugs treating similar medical conditions often share similar side effect patterns due to comparable mechanisms of action.
 - This insight highlights the importance of monitoring side effects when multiple treatment options are available.
-

7.5 Prescription vs Over-the-Counter (Rx vs OTC) Insights

- Prescription (Rx) drugs generally have higher review counts, indicating more detailed patient feedback.
 - OTC drugs tend to have slightly lower average ratings but are more accessible to users.
 - Rx/OTC classification plays an important role in understanding drug usage patterns and safety requirements.
-

7.6 Risk and Regulation Insights

- Drugs classified under higher **CSA schedules** are fewer but require careful regulation.
 - Pregnancy categories reveal that many drugs fall under **Category C**, indicating potential risks that must be weighed against benefits.
 - Alcohol interaction data shows that some drugs should be used with caution due to possible adverse interactions.
-

7.7 Overall Observations

- Drugs with higher reviews generally have more stable and reliable ratings.
- Medical conditions with more treatment options show greater variability in drug ratings and side effects.
- Combining ratings, reviews, and safety indicators provides a more comprehensive understanding of drug effectiveness.

8.2 Future Scope

The project can be extended and improved in several ways:

1. Advanced Side Effect Analysis

- Apply text mining or NLP techniques to group similar side effects.
- Identify severe vs mild side effects automatically.

2. Sentiment Analysis on Reviews

- Analyze user review text to understand emotional sentiment.
- Compare sentiment scores with numeric ratings.

3. Time-Based Trends

- Study how ratings and reviews change over time.
- Identify drugs with improving or declining user satisfaction.

4. Predictive Modeling

- Build machine learning models to predict drug ratings based on side effects and medical conditions.
- Predict potential adverse reactions for new drugs.

5. Enhanced Dashboard Features

- Add drill-through pages for individual drugs or medical conditions.
- Include dynamic tooltips with safety warnings and interaction alerts.

6. Healthcare Decision Support

- Use insights to support doctors, pharmacists, and patients in selecting safer and more effective medications.